

Verfahren und System zum Erfassen von Daten aus maschinell lesbaren Dokumenten

- 5 Die Erfindung betrifft ein Verfahren und ein System zum Erfassen von Daten aus maschinell lesbaren Dokumenten, wobei die Daten einer Datenbank zugeordnet werden, in dem einzelne Daten möglichst automatisch dem Dokument extrahiert werden und in entsprechende Datenbankfelder eingetragen werden, wo-
- 10 bei das erfindungsgemäße Verfahren und System das Erfassen von Daten betrifft, falls Daten für ein oder mehrere bestimmte Datenbankfelder eines Dokumentes nicht mit der notwendigen Zuverlässigkeit extrahiert werden konnten.
- 15 Verfahren und Systeme zum Erfassen von Daten aus maschinell lesbaren Dokumenten sind bekannt. Üblicherweise weisen die Systeme einen Scanner auf, mit welchem Vorlagen optisch abgetastet werden. Die hierbei erzeugten Dateien sind maschinell lesbare Dokumente und enthalten in der Regel Textelemente.
- 20 Die Textelemente werden mit Hilfe einer OCR-Einrichtung in codierten Text umgesetzt. Den Dateien werden in der Regel vorbestimmte Formulare bzw. Templates zugeordnet, so dass anhand der Formulare gezielt bestimmte Informationen aus den Text enthaltenden Dateien ermittelt werden können. Diese In-
- 25 formationen werden zum Beispiel in einer Datenbank abgespeichert.

Derartige Verfahren und Systeme werden beispielsweise bei großen Firmen eingesetzt, um Rechnungen zu lesen. Die so extrahierten Daten können automatisch einer betriebswirtschaftlichen Software übermittelt werden.

30

Ein solches System ist in der US 4,933,979 beschrieben. Dieses System weist einen Scanner zum optischen Abtasten von Formularen auf. Bei diesem System können eine Vielzahl von Formulartypen definiert werden, wobei jeder Formulartyp bzw. Template durch mehrere Parameter, insbesondere geometrisch

35

BEST AVAILABLE COPY

definierte Bereiche, in welchen Texte oder Bilder enthalten sein sollen, festgelegt ist. Die Formulartypen können auch durch weitere Eigenschaften, wie zum Beispiel der Schrift, die in den Texten enthalten ist (Alphabet, Zahlen, Symbole, Katakana, Kanji, Handschrift) definiert sein. Nach dem Scannen eines Formulars wird mittels einer Formulartypunterscheidungseinrichtung dem gescannten Formular ein Template zugeordnet. Dementsprechend werden die in dem Textfeld enthaltenen Daten mittels einer OCR-Einrichtung gelesen und extrahiert. Falls kein geeignetes Template vorhanden ist, muss eines erstellt werden.

Aus der WO 98/47098 geht ein weiteres System zum automatischen Erfassen von Daten aus maschinell lesbaren Dokumenten hervor. Hierbei werden mittels eines Scanners Formulare optisch abgetastet. Danach wird automatisch eine Linien-Karte des Formulars erstellt. Hierbei werden zum einen alle Linien erfasst als auch grafische Elemente in eine Linienstruktur umgesetzt. Andere Elemente, wie zum Beispiel Textabschnitte, werden ausgefiltert. Alle vertikalen Linien bilden die Grundlage zur Erstellung eines vertikalen Schlüssels und alle horizontalen Linien bilden die Grundlage zur Erstellung eines horizontalen Schlüssels. Danach wird ermittelt, ob bereits ein Template mit einem korrespondierenden vertikalen und horizontalen Schlüssel vorhanden ist. Falls dies der Fall ist, werden die Daten mit einem entsprechenden Template ausgelesen. Ist dies nicht der Fall, so wird anhand des eingescannten Formulars mittels eines Selbstlern-Modus ein Template erstellt und abgespeichert.

30

In dem Buch Modern Information Retrieval von Baeza-Yates und Ribeiro-Neto, Eddison-Wessley Verlag, ISBN 0-201-39829-X sind die Grundlagen von Datenbanken und zum schnellen Wiederauffinden von in Datenbanken gespeicherten Informationen erläutert. So ist im Kapitel 8.2 ein Verfahren mit invertierten Dateien, das auch als invertierter Index bezeichnet wird, beschrieben. Bei diesem Verfahren wird aus einem zu untersu-

35

chenden Text zunächst ein Wörterbuch mit allen im Text enthaltenen Wörtern erstellt. Allen Wörtern des Wörterbuches werden eine oder mehrere Zahlen zugeordnet, die angeben, an welcher Stelle das Wort im Text auftritt. Derartige invertierte Dateien erlauben eine schnellere automatische Analyse eines zu durchsuchenden Textes. Im Kapitel 8.6.1 ist ein String Matching-Verfahren beschrieben, mit welchem zwei Strings verglichen werden und ein zur Ähnlichkeit der Strings indirekt proportionales Kostenmaß berechnet wird. Wenn die beiden Strings identisch sind, ist der Betrag des Kostenmaßes Null. Je stärker sich die Strings unterscheiden, desto größer ist der Betrag des Kostenmaßes. Das Kostenmaß ist somit ein Ausdruck für die Ähnlichkeit der beiden Strings. Dieses und ähnliche Verfahren sind auch unter den Bezeichnungen Approximate String Matching, Levenshtein-Verfahren, Elastic Matching und Viterbi-Algorithmus bekannt. Diese Verfahren gehören zu dem Gebiet der dynamischen Programmierung.

Aus der noch nicht veröffentlichten Patentanmeldung DE 103 42 594.2 geht ein Verfahren und ein System zum Erfassen von Daten aus mehreren maschinell lesbaren Dokumenten hervor, bei dem aus einem zu bearbeitenden Dokument, dem Lesedokument, Daten extrahiert werden, indem sie an Positionen aus dem Lesedokument ausgelesen werden, die durch in einem Vorlagedokument eingetragene Felder bestimmt sind.

Tritt ein Fehler beim Auslesen der Lesedokumente auf, wird das Lesedokument an einem Bildschirm dargestellt und lediglich durch Markieren entsprechender Felder im Lesedokument können die Daten ausgelesen werden. Hierbei werden, falls es erforderlich ist, automatisch weitere Vorlagedokumente anhand der markierten Lesedokumente erstellt bzw. vorhandene Vorlagedokumente entsprechend korrigiert. Dieses System ist derart einfach bedienbar, so dass keine speziellen Computer- oder Softwarekenntnisse notwendig sind.

Der Erfindung liegt die Aufgabe zugrunde, ein Verfahren und ein System zum Erfassen von Daten aus maschinell lesbaren Dokumenten zu schaffen, bei dem die Eingabe der Daten gegenüber dem bekannten Verfahren erheblich vereinfacht wird, falls Daten nicht automatisch extrahiert werden konnten.

Die Aufgabe wird durch ein Verfahren mit den Merkmalen des Anspruchs 1 und durch ein System mit dem Merkmal des Anspruchs 16 gelöst. Vorteilhafte Ausgestaltungen der Erfindung sind in den jeweiligen Unteransprüchen angegeben.

Mit den oben erläuterten Verfahren können Daten aus mehreren maschinell lesbaren Dokumenten erfasst werden, wobei die Daten einer Datenbank zugeordnet werden, indem einzelne Daten möglichst automatisch dem Dokument extrahiert werden und in entsprechende Datenbankfelder eingetragen werden. Falls Daten für ein oder mehrere bestimmte Datenbankfelder einem Dokument nicht mit der notwendigen Zuverlässigkeit extrahiert werden konnten, zum Beispiel weil ein Fehler festgestellt worden ist, der beispielsweise dadurch verursacht sein kann, dass in dem Dokument an der Stelle, wo die Daten gelesen werden sollten, keine Daten oder falsche Daten vorhanden sind, oder dass beim Einlesen dieses Dokumentes mittels eines OCR-Verfahrens ein oder mehrere Zeichen falsch umgesetzt werden, so werden erfindungsgemäß folgende Schritte ausgeführt:

- Darstellen des Dokumentes an einem Bildschirm,
- Anzeigen des Datenbankfeldes, für das die Daten nicht mit der notwendigen Zuverlässigkeit extrahiert werden konnten, am Bildschirm
- Ausführen einer Vorschlags-Routine, mit welcher Stringabschnitte in der Nähe eines von einem Benutzer auf dem Bildschirm bewegbaren Zeiger ausgewählt, markiert und zur Extraktion vorgeschlagen werden.

Das Dokument wird am Bildschirm dargestellt, damit der Benutzer es lesen kann. Zudem wird das Datenbankfeld angezeigt,

für das die Daten nicht mit der notwendigen Zuverlässigkeit extrahiert werden konnten. Hierdurch wird der Benutzer in Kenntnis darüber gesetzt, für welches Datenbankfeld die Daten aus dem am Bildschirm dargestellten Dokument noch zu extrahieren sind.

Durch das Ausführen bzw. Aktivieren der Vorschlags-Routine werden Stringabschnitte in der Nähe eines von dem Benutzer auf dem Bildschirm bewegbaren Zeiger ausgewählt, markiert und zur Extraktion vorgeschlagen. Hierdurch muss der Benutzer lediglich den Zeiger auf dem am Bildschirm dargestellten Dokument in die Nähe eines Stringabschnittes bewegen, der die Daten für das angezeigte Datenbankfeld enthält. Die Daten werden dann automatisch ausgewählt, markiert und zur Extraktion vorgeschlagen. Der Benutzer kann dann lediglich durch Betätigen einer bestimmten Taste den vorgeschlagenen Stringabschnitt in das Datenbankfeld übernehmen.

Durch das automatische Auswählen und Markieren des Stringabschnittes wird der Vorgang der Übernahme der noch fehlenden Daten erheblich vereinfacht und beschleunigt.

Nach einer bevorzugten Ausführungsform der Erfindung werden beim Auswählen des Stringabschnittes Konzept-Informationen berücksichtigt, die dem jeweiligen Datenbankfeld zugeordnet sind.

Die Erfindung wird nachfolgend beispielhaft anhand der Zeichnung näher erläutert. In der Zeichnung zeigen:

- Figur 1 ein Verfahren zum Erfassen von Daten aus einem Dokument, die nicht automatisch extrahiert werden konnten,
- Figur 2 - 6 jeweils Kopien von Bildschirmdarstellungen zu einzelnen Verfahrensschritten des in Figur 1 gezeigten Verfahrens,

Figur 7 ein Verfahren zum Extrahieren von in Tabellen angeordneten Daten,

5 Figur 8, 9 jeweils eine Tabelle mit markierten Daten, und

Figur 10 ein System zum Ausführen des erfindungsgemäßen Verfahrens.

10 Das erfindungsgemäße Verfahren zum Erfassen von Daten aus maschinell lesbaren Dokumenten ist eine Weiterbildung der eingangs erläuterten Verfahren, mit welchen aus Dokumenten maschinell Daten extrahiert und in einer Datenbank gespeichert werden können.

15 Bei diesen Verfahren können jedoch nicht immer alle Datenbankfelder der Datenbank zuverlässig mit aus den Dokumenten extrahierten Daten gefüllt werden. Liegt zum Beispiel beim Extrahieren der Daten ein Fehler vor, so wird das automatische Verfahren unterbrochen und unter Mitwirkung eines Benutzers werden die Daten aus dem Dokument manuell in Datenbankfelder übertragen. Ein solcher Fehler kann dadurch verursacht sein, dass in dem zu bearbeitenden Dokument kein geeigneter Stringabschnitt gefunden wird, aus dem die Daten gelesen werden können oder der Stringabschnitt fehlerhafte Daten enthält, die beispielsweise beim Umsetzen in codierten Text des Dokumentes mittels eines OCR-Verfahrens entstanden sind.

20

25

Das erfindungsgemäße Verfahren beginnt somit dann, wenn Daten nicht zuverlässig extrahiert werden können. Der Ausdruck „nicht zuverlässig extrahierbar“ umfasst sowohl grundsätzliche Fehler beim Lesen von Daten, die ein Lesen der Daten nicht möglich machen als auch gelesene Daten, die zum Beispiel unter Berücksichtigung von Kontext-Informationen auf das Datenbankfeld abgebildet werden, wobei die Güte der Abbildung ermittelt wird. Derartige Abbildungsverfahren sind zum Beispiel das eingangs erläuterte String Matching-

30

35

Verfahren. Ist die hierbei erzielte Abbildungsgüte zu gering, so werden die automatisch eingelesenen Daten als nicht ausreichend zuverlässig bewertet und verworfen.

5 Nachfolgend wird das erfindungsgemäße Verfahren anhand des in Figur 1 dargestellten Flussdiagramms erläutert. In dem Flussdiagramm sind alle Schritte, die automatisch ausgeführt werden, mit einem „a“ im Kreis und alle vom Benutzer manuell zu tätigen Schritte mit einem „m“ im Kreis gekennzeichnet.

10

Es beginnt mit dem Schritt S1.

Nachdem zumindest Daten für ein Datenbankfeld nicht mit der notwendigen Zuverlässigkeit extrahiert werden konnten, wird  
15 das entsprechende Dokument 1 an einem Bildschirm 2 dargestellt und das Datenbankfeld 3 angezeigt (Schritt S2). Figur 2 zeigt eine Bildschirmdarstellung unmittelbar nach dem Feststellen, dass Daten nicht mit der notwendigen Zuverlässigkeit extrahiert werden konnten, wobei in einem Fenster 4/1 auf der  
20 rechten Seite der Bildschirmdarstellung das Dokument 1 dargestellt ist. Auf der linken Seite sind zwei Fenster 4/2 und 4/3 angeordnet. Das Fenster 4/2 enthält eine Übersicht der zu bearbeitenden Dokumente und im Fenster 4/3 sind die einzelnen Datenbankfelder angegeben, in welche Daten gespeichert werden,  
25 den, die aus dem Dokument 1 zu lesen sind.

In dem dargestellten Beispiel konnte keines der Datenbankfelder mit Daten gefüllt werden, weshalb die einzelnen Datenbankfelder 3 mit dem Zusatz [empty] versehen sind. Es ist jedoch auch möglich, dass nur in wenigen Datenbankfeldern oder  
30 lediglich in einem einzigen Datenbankfeld Daten fehlen.

In Figur 2 ist das Datenbankfeld „InvoiceNumber“ (= Rechnungsnummer) im Vergleich zu den anderen Datenbankfeldern 3  
35 dunkler markiert, was dem Benutzer anzeigt, dass für dieses Datenbankfeld 3 Daten aus dem Dokument 1 zu extrahieren sind. Zusätzlich ist im oberen Bereich des Fensters 4/1 in großer

Schrift der Begriff „InvoiceNumber“ aufgeführt, das dem Benutzer zusätzlich anzeigt, für welches Datenbankfeld Daten zu extrahieren sind.

- 5 Der Benutzer kann nun im Fenster 4/1 einen Zeiger 5 positionieren, den er vorzugsweise derart anordnet, dass er sich möglichst nahe an dem Stringabschnitt befindet, von dem der Benutzer annimmt, dass dessen Inhalt in dem entsprechenden Datenbankfeld abzuspeichern ist. In dem in Figur 2 gezeigten
- 10 Beispiel sind Daten für das Datenbankfeld „Rechnungsnummer“ zu extrahieren, weshalb der Zeiger 5 in der Nähe der Rechnungsnummer „4361“ positioniert wird (Schritt S3).

- Der Zeiger 5 kann hierbei mittels einer Maus 6 oder durch
- 15 Eingabe an einer Tastatur 7 im Fenster 4/1 bewegt werden.

- Nach dem Positionieren des Zeigers 5 beginnt eine Vorschlags-Routine, die mehrere Verfahrensschritte umfasst. Diese Vorschlags-Routine kann einerseits dadurch ausgelöst werden, dass der Zeiger 5 ein vorbestimmtes Zeitintervall nicht
- 20 bewegt wird, wodurch dann die Vorschlags-Routine automatisch ausgeführt wird, oder dadurch, dass eine bestimmte Taste einer Maus oder der Tastatur bestätigt wird.

- 25 Im Schritt S4 wird zunächst geprüft, ob in der näheren Umgebung des Zeigers ein Stringabschnitt mit einem für das Datenbankfeld 3 geeigneten Konzept vorhanden ist, sofern dem entsprechenden Typ des Datenbankfeldes vorab Konzeptinformationen zugeordnet wurden. Diese Konzeptinformationen umfassen
- 30 die Syntax und/oder die Semantik des Datenbankfeldes. Informationen zur Syntax sind zum Beispiel die Anzahl von Ziffern und/oder Buchstaben und/oder vorbestimmte Formate des zu lesenden Stringabschnittes. So weisen Datumsfelder, Betragsfelder und Adressfelder in der Regel bestimmte Formate auf. In-
- 35 formationen zur Semantik umfassen vorbestimmte Begriffe, die in das entsprechende Datenbankfeld eingefügt werden können. Dies ist zum Beispiel bei Währungsangaben zweckmäßig, oder



wenn die Artikelbezeichnung eines bestimmten Lieferanten eingelesen werden sollen, der eine begrenzte Anzahl von Artikeln liefern kann. Die entsprechenden Artikelbezeichnungen sind dann in einem Lexikon abgelegt und können dann eindeutig erkannt werden.

Bei dem in Figur 2 dargestellten Ausführungsbeispiel befinden sich in der Nähe des Zeigers 5 die zwei Stringabschnitte „4361“ und „02.08.2002“. Der letzte Stringabschnitt besitzt die Syntax eines Datums, weshalb er zum Extrahieren der Rechnungsnummer verworfen wird. Der Stringabschnitt „4361“ entspricht der Syntax einer Rechnungsnummer. Somit wird im Schritt S4 entschieden, dass ein Stringabschnitt mit einem geeigneten Konzept vorliegt, weshalb der Verfahrensablauf auf den Schritt S5 übergeht. Im Schritt S5 wird der Stringabschnitt „4361“ markiert (Figur 3). Die Markierung erfolgt im vorliegenden Ausführungsbeispiel durch eine farbliche Unterlegung des Stringabschnittes und durch Zeichnen eines Rahmens 8.

Sollte im Schritt S4 kein geeignetes Konzept ermittelt werden, so geht der Verfahrensablauf auf den Schritt S6 über. Im Schritt S6 wird das zum Zeiger 5 nächstliegend angeordnete Einzelzeichen ermittelt, das im vorliegenden Ausführungsbeispiel gemäß Fig. 2 - 4 die „1“ ist. Danach werden die Grenzen des dieses Zeichen enthaltenden Stringabschnittes nach allgemeinen Regeln ermittelt. Diese Grenzen können zum Beispiel durch Leerzeichen bzw. Leerräume im Dokument 1 oder durch bestimmte Satzzeichen oder sonstige Markierungen im Dokument 1 vorgegeben sein. Werden entsprechende Begrenzungsmarkierungen erkannt, so wird der dazwischen liegende Stringabschnitt ausgewählt und markiert. Bei dem in Figur 2 und 3 gezeigten Ausführungsbeispiel befinden sich seitlich des Stringabschnittes „4361“ jeweils Leerräume, durch die auch nach den allgemeinen Regeln eine eindeutige Auswahl der Markierung des Stringabschnittes möglich ist.

Unabhängig davon, ob der Stringabschnitt gemäß dem Schritt S5 oder gemäß dem Schritt S6 ausgewählt oder markiert worden ist, geht der Verfahrensablauf auf den Verfahrensschritt S7 über, mit der der Stringabschnitt in einem zusätzlichen Rahmen 9 als codierter Text und in einem weiteren Rahmen 10 vergrößert dargestellt wird (Fig. 3, 4). Bei dem vorliegenden Ausführungsbeispiel liegt das Dokument 1 als grafische Datei, zum Beispiel im pdf, tif, gif, jpg-Format vor. Üblicherweise wird in dem bereits vorausgegangenen Verfahrensabschnitt das Dokument einer OCR-Routine unterzogen und in codierten Text umgesetzt. Der codierte Text ist hierbei auch auf Konzepte untersucht worden und die entsprechenden Informationen sind abgespeichert worden. Diesem codierten Text wird der zu dem Stringabschnitt korrespondierende Abschnitt entnommen und in dem Rahmen 9 dargestellt. Hierdurch erkennt der Benutzer, ob der Stringabschnitt korrekt in codierten Text umgesetzt worden ist.

Im Rahmen 10 wird der Stringabschnitt im Grafikformat in vergrößerter Darstellung dargestellt, wodurch der Benutzer auch Details im Stringabschnitt erkennen kann.

Mit dem Schritt S7 ist die Vorschlags-Routine abgeschlossen.

Im Schritt S8 beurteilt der Benutzer, ob der ausgewählte und markierte Stringabschnitt grundsätzlich zur Übernahme in das Datenbankfeld geeignet ist. Ist dies nicht der Fall, so wird der Zeiger 5 erneut positioniert (S3) und die Vorschlags-Routine (S4 - S7) wiederholt ausgeführt. Ist die Auswahl des Stringabschnitts hingegen grundsätzlich geeignet, so beurteilt der Benutzer, ob auch der markierte Bereich korrekt ist (Schritt S9). Ist dies nicht der Fall, so kann der Benutzer die Markierung des Stringabschnittes manuell bearbeiten und/oder den codierten Text im Rahmen 9 editieren (Schritt S10). Mit dem Editieren des codierten Textes können Fehler, die durch eine nicht korrekte OCR-Umsetzung entstanden sind, behoben werden. Bei diesen Korrekturen (Bereich anpassen, E-

ditieren) werden automatisch der markierte Bereich und die Inhalte der Rahmen 9 und 10 angepasst.

Ist der markierte Bereich korrekt bzw. vom Benutzer entsprechend überarbeitet worden, so geht der Verfahrensablauf auf den Schritt S11 über, mit dem die in dem ausgewählten Stringabschnitt enthaltenen Daten in das korrespondierende Datenbankfeld übertragen werden (Figur 4). Diese Übertragung der Daten wird durch eine Betätigung einer vorbestimmten Taste an der Maus oder der Tastatur vom Benutzer ausgelöst. Danach ist das Verfahren zum Extrahieren von Daten für ein Datenbankfeld beendet (S12). Sind Daten für weitere Datenbankfelder zu lesen, so beginnt das Verfahren erneut mit dem Schritt S1. In Figur 5 ist das nächste zu lesende Datenbankfeld „Invoice Date“ (= Rechnungsdatum) angezeigt.

Mit dem erfindungsgemäßen Verfahren wird die Tätigkeit eines Benutzers beim manuellen Übertragen von Daten aus einem Dokument in ein Datenbankfeld lediglich auf das Positionieren des Zeigers, der Kontrolle der automatisch vorgeschlagenen Auswahl und der eventuellen Korrektur des Bereiches und das Betätigen einer Taste zum Übertragen der Daten beschränkt. Die Auswahl und die Markierung des Bereichs des auszuwählenden Stringabschnittes wird vom erfindungsgemäßen Verfahren selbsttätig ausgeführt.

Die Figuren 2 bis 5 zeigen die Übernahme von Daten in ein einzelnes Datenbankfeld. Durch die Berücksichtigung von Konzept-Informationen ist es jedoch auch möglich, mit einem einzigen Stringabschnitt Daten für mehrere Datenbankfelder zu extrahieren. Figur 6 zeigt ein entsprechendes Ausführungsbeispiel, in dem die vollständige Adresse als ein Stringabschnitt markiert und gelesen wird, wobei die Adresse selbsttätig in die einzelnen Datenbankfelder, Name, Firma, Straße, Postleitzahl und Stadt segmentiert wird.

Nachfolgend wird eine weitere Ausgestaltung des oben beschriebenen Verfahrens anhand des Flussdiagramms aus Figur 7 und der Bildschirmdarstellungen gemäß Figur 8 und 9 erläutert, mit dem Daten aus Tabellen extrahiert werden können.

5

Dieses Verfahren beginnt mit dem Schritt S15.

Im Schritt S16 werden die Werte der Tabelle in der ersten Tabellenzeile gemäß obigen Verfahren durch Positionieren des  
10 Zeigers, automatisches Auswählen und Markieren des Stringabschnittes und Übernehmen der Daten in korrespondierende Datenbankfelder extrahiert. Figur 8 zeigt eine Tabelle, in der die Stringabschnitte der ersten Tabellenzeile markiert sind, die in die entsprechenden Datenbankfelder übernommen worden  
15 sind. Diese Datenbankfelder besitzen die Struktur einer Tabelle, zum Beispiel sind sie als zwei-dimensionales Datenfeld angelegt, so dass beim Extrahieren der Daten in diese Datenbankfelder das Verfahren anhand des Datenbankfeldes selbstständig erkennt, dass Daten aus einer Tabelle ausgelesen werden.

20

Eine Tabellenzeile kann sich auch über mehrere Seiten erstrecken, wenn die Tabelle sich entsprechend über mehrere Seiten erstreckt. Sind die Daten der ersten Tabellenzeile vollständig extrahiert worden, kann der Benutzer durch eine  
25 vorbestimmte Eingabe das automatische Extrahieren der weiteren Tabelleneinträge initiieren. Wird diese Eingabe vom Benutzer getätigt, so werden im Schritt S17 zunächst eine Liste mit allen Stringabschnitten erstellt, die unterhalb der ersten Tabellenzeile angeordnet sind.

30

In Schritt S18 wird mittels einer Kostenfunktion ein Kostenwert zwischen Sequenzen von Stringabschnitten der Liste und der Sequenz der Stringabschnitte der ersten Tabellenzeile, auf deren Grundlage Daten in die Datenbankfelder im Schritt  
35 S16 extrahiert worden sind, ermittelt. Bei dieser Kostenfunktion werden den Sequenzen der Stringabschnitte der Liste geringe Kosten zugewiesen, deren Stringabschnitte bezüglich ih-

rer horizontalen Position und ihrer Breite mit den korrespondierenden Stringabschnitten der ersten Tabellenzeile übereinstimmen oder zumindest sehr ähnlich sind. Dieser Kostenwert ist somit indirekt proportional zur Ähnlichkeit zwischen den  
5 in der Liste aufgeführten Sequenzen von Stringabschnitten und den in der ersten Tabellenzeile enthaltenen Sequenz von Stringabschnitten.

Die hierbei verwendete Kostenfunktion entspricht der aus dem  
10 Kapitel 8.6.1 String Matching Allowing Errors in Modern Information Retrieval (ISBN 0-201-39829-X) beschriebenen Kostenfunktion, mit welcher jeweils ein Einzelkostenwert zwischen einem Stringabschnitt der ersten Tabellenzeile und einem Stringabschnitt der weiteren Tabellenzeilen ermittelt  
15 wird. Da jede Sequenz mehrere Stringabschnitte umfasst, wird mittels des Viterbi-Algorithmus für die einzelnen Sequenzen von Stringabschnitten jeweils ein Gesamtkostenwert bzw. Gesamtähnlichkeitswert durch Summieren der Einzelkostenwerte berechnet.

20 Anhand dieser Kostenwerte bzw. Ähnlichkeitswerte werden die Sequenzen von Stringabschnitten als Tabellenzeilen bestimmt, deren Ähnlichkeitswert unter einem vorbestimmten Schwellwert liegt (S19). Hierdurch sind alle Tabellenzeilen und damit Tabelleneinträge der Tabelle bestimmt. Sie werden im Schritt  
25 S20 markiert (Figur 9) und im Schritt S21 extrahiert, d.h., automatisch ausgelesen, gegebenenfalls in codierten Text umgesetzt, und in den entsprechenden Datenbankfeldern gespeichert.

30 Mit dem Schritt S22 ist dieses Verfahren beendet.

Zweckmäßigerweise ist es möglich, die Tabelleneinträge nachzubearbeiten, d.h., die markierten Bereiche zu verändern  
35 (verschieben, vergrößern, verkleinern) bzw. einzelne Zeilen zu entfernen bzw. hinzuzufügen. Bei einer Nachbearbeitung

werden die Einträge in den Datenbankfeldern automatisch entsprechend angepasst.

5 Zusätzlich kann beim Auslesen der Daten und Eintragen in die Datenbankfelder eine zusätzliche Kontrolle durch eine Abbildung mittels dem String Matching-Verfahren erfolgen, mit welcher bestimmt wird, wie gut die Einträge mit dem durch die einzelnen Datenbankfelder vorgegebenen Konzept übereinstimmen.

10

Weiterhin kann das erfindungsgemäße Verfahren mit dem in der deutschen Patentanmeldung DE 103 42 594.2 beschriebenen Verfahren zum Erfassen von Daten aus mehreren maschinell lesbaren Dokumenten kombiniert werden, weshalb auf die Patentanmeldung vollinhaltlich Bezug genommen wird und sie durch Bezugnahme in die vorliegende Patentanmeldung inkorporiert wird.

20 Bei diesem Verfahren zum automatischen Erfassen von Daten aus mehreren maschinell lesbaren Dokumenten werden Vorlagedokumente mit einem Lesedokument verglichen und deren Ähnlichkeit bewertet. Das hierbei angewandte Verfahren kann auch zum Auslesen einer Tabelle angewendet werden, wobei die Sequenz der ausgewählten Stringabschnitte der ersten Tabellenzeile dem Vorlagedokument entsprechen und die Kombinationen von  
25 Stringabschnitten der weiteren Tabellenzeilen den Lesedokumenten entsprechen.

Bei dem oben beschriebenen erfindungsgemäßen Verfahren zum  
30 Extrahieren von Daten aus Tabellen muss ein Benutzer lediglich den Zeiger zu den Tabelleneinträgen in der ersten Tabellenzeile bewegen und die Übernahme der dann automatisch ausgewählten und markierten Stringabschnitte als Daten für das entsprechende Datenbankfeld bestätigen. Nachdem er dies für  
35 alle Tabelleneinträge der ersten Tabellenzeile ausgeführt hat, muss er lediglich durch eine Eingabe das vollständige Auslesen der weiteren Tabelleneinträge initiieren. Das Ver-

fahren ermittelt dann selbstständig die weiteren Tabelleneinträge, markiert sie und extrahiert die Daten in die Datenbank.

- 5 Dies beschleunigt erheblich das Auslesen von Daten aus der Tabelle in eine Datenbank. Der Verfahrensabschnitt S17 bis S21 stellt daher eine eigenständige Erfindung dar, die jedoch bevorzugt in Kombination mit dem in Figur 1 dargestellten Verfahren, auf der sich der Schritt S16 bezieht, verwendet  
10 wird.

Figur 10 zeigt schematisch ein System zum Ausführen des erfindungsgemäßen Verfahrens. Dieses System 11 weist einen Computer 12 mit einer Speichereinrichtung 13, mit einer zentralen  
15 Prozessoreinrichtung (CPU) 14 und eine Interface-Einrichtung 15 auf. Am Computer 12 sind ein Scanner 16, ein Bildschirm 2 und eine Eingabeeinrichtung 17 angeschlossen. Die Eingabeeinrichtung 17 umfasst eine Tastatur 7 und/oder eine Maus 6.

- 20 In der Speichereinrichtung 13 ist ein Softwareprodukt zum Ausführen des erfindungsgemäßen Verfahrens gespeichert, das an der CPU 14 ausgeführt wird. Mit dem Scanner 16 werden Dokumente erfasst und in eine elektronische Datei umgewandelt.  
25 Diese elektronischen Dateien werden vom Computer 12 eingelesen und eventuell mittels einer OCR-Routine und/oder eines Verfahrens zum Erkennen spezieller Syntax oder Semantik in der Datei vorverarbeitet. Danach werden die in den Dateien enthaltenen Dokumente entsprechen dem oben beschriebenen Verfahren mit dem System 11 bearbeitet. An der Eingabeeinrich-  
30 tung 17 können die entsprechenden Eingaben vorgenommen werden, wobei diese sich lediglich auf Bewegungen des Zeigers 5 und ein paar wenige Tastatureingaben beschränken. Gegebenenfalls werden die markierten Felder mittels Tastatur oder Mauseingabe  
35 verschoben oder durch Vergrößerung bzw. Verkleinerung angepasst oder der codierte Text editiert.

Die Erfindung ist oben anhand eines Ausführungsbeispiels erläutert worden. Im Rahmen der Erfindung sind hiervon Abwandlungen möglich. So kann zum Beispiel anstelle des Rahmens 8 lediglich der Rahmen 10 vorgesehen werden, in dem der ausgewählte Stringabschnitt vergrößert dargestellt wird. Dieser Rahmen 10 stellt auch eine Markierung des Stringabschnittes dar.

Bei dem oben erläuterten Ausführungsbeispiel werden die Dokumente eingescannt und liegen dann in einem Graphikformat vor. Das erfindungsgemäße Verfahren kann jedoch auch zum Lesen von Informationen aus Dokumenten angewandt werden, die bereits in codiertem Text vorliegen, wie zum Beispiel E-Mails. Bei einer solchen Anwendung ist es selbstverständlich nicht notwendig, dass die Dokumente mittels einer OCR-Routine in codierten Text umgesetzt werden.

Die Erfindung kann folgendermaßen kurz zusammengefasst werden:

20

Die Erfindung betrifft ein Verfahren zum Erfassen von Daten aus maschinell lesbaren Dokumenten, wobei die Daten einer Datenbank zugeordnet werden.

25 Mit der Erfindung werden Stringabschnitte, die sich in der Nähe eines vom Benutzer bewegbaren Zeigers befinden, automatisch ausgewählt, markiert und deren Inhalt zur Übernahme in eine Datenbank vorgeschlagen.

30 Nach einer Weiterbildung des erfindungsgemäßen Verfahrens kann der Inhalt einer Tabelle vollautomatisch ausgelesen werden, wenn die Tabelleneinträge in einer ersten Tabellenzeile einmal gemäß obigen Verfahren ausgelesen worden sind.

35 Es wurden Ausführungsbeispiele der Erfindung beschrieben. Dabei ist klar, dass der Fachmann jederzeit Abwandlungen und Weiterbildungen angeben kann, die das erfindungsgemäße Kon-



zept benutzten. Weiterhin kann die Erfindung sowohl mittels elektronischen Komponenten (Hardware) als auch durch Computerprogrammelemente (Software oder Softwaremodule) realisiert werden. Die Erfindung wird dabei insbesondere aus einer Kombination von elektronischen Hardware-Elementen und Software-  
5 elementen realisiert. Dementsprechend erfaßt die Erfindung auch Computerprogrammprodukte wie z.B. elektronische Datenträger (CD, DVD, Diskette, Bandlaufwerke) bzw. Komponenten, die über Computernetzwerke (Internet) verbreitet werden  
10 und/oder auf Computern und insbesondere im Zwischenspeichern geladen, bereit gehalten und/oder zum Ablauf gebracht werden.

Bezugszeichenliste

- 1 Dokument
- 2 Bildschirm
- 5 3 Datenbankfeld
- 4 Fenster
- 5 Zeiger
- 6 Maus
- 7 Tastatur
- 10 8 Rahmen
- 9 Rahmen
- 10 Rahmen
- 11 System
- 12 Computer
- 15 13 Speichereinrichtung
- 14 CPU
- 15 Interface-Einrichtung
- 16 Scanner
- 17 Eingabeeinrichtung

## Ansprüche

1. Verfahren zum Erfassen von Daten aus maschinell lesbaren  
5 Dokumenten, wobei die Daten einer Datenbank zugeordnet werden, indem einzelne Daten möglichst automatisch dem Dokument extrahiert werden und in entsprechende Datenbankfelder eingetragen werden, und falls Daten für ein  
10 oder mehrere bestimmte Datenbankfelder einem Dokument nicht mit der notwendigen Zuverlässigkeit extrahiert werden konnten, werden folgende Schritte ausgeführt:
  - Darstellen des Dokumentes an einem Bildschirm,
  - Anzeigen des Datenbankfeldes, für das die Daten nicht  
15 mit der notwendigen Zuverlässigkeit extrahiert werden konnten, am Bildschirm,
  - Ausführen einer Vorschlags-Routine, mit welcher Stringabschnitte in der Nähe eines von einem Benutzer auf dem  
20 Bildschirm bewegbaren Zeiger ausgewählt, markiert und zur Extraktion vorgeschlagen werden.
2. Verfahren nach Anspruch 1,  
d a d u r c h g e k e n n z e i c h n e t,  
dass der Stringabschnitt nach Maßgabe von dem Datenbank-  
25 feld zugeordneten Konzept-Informationen ausgewählt, markiert und zur Extraktion vorgeschlagen wird.
3. Verfahren nach Anspruch 2,  
d a d u r c h g e k e n n z e i c h n e t,  
dass die Konzept-Informationen die Syntax und/oder die  
30 Semantik des Datenbankfeldes beschreiben, so dass von der Vorschlags-Routine ein zu markierender Stringabschnitt entsprechend der Syntax bzw. der Semantik des jeweiligen  
Datenbankfeldes ausgewählt und markiert wird.

4. Verfahren nach Anspruch 3,  
d a d u r c h g e k e n n z e i c h n e t,  
dass die Informationen zur Syntax die Anzahl von Ziffern  
und/oder Buchstaben und/oder vorbestimmte Formate des zu  
5 lesenden Stringabschnitts beschreiben.
5. Verfahren nach Anspruch 3 oder 4,  
d a d u r c h g e k e n n z e i c h n e t,  
dass die Informationen zur Semantik vorbestimmte Begriffe,  
10 bspw. mit einem Lexikon, beschreiben.
6. Verfahren nach einem der Ansprüche 1 bis 5,  
d a d u r c h g e k e n n z e i c h n e t,  
dass ein Stringabschnitt ausgewählt wird, der zwischen  
15 zwei Begrenzungszeichen angeordnet ist.
7. Verfahren nach Anspruch 6,,  
d a d u r c h g e k e n n z e i c h n e t,  
dass Begrenzungszeichen Leerzeichen und/oder Satzzeichen  
20 umfassen.
8. Verfahren nach einem der Ansprüche 1 bis 7,  
d a d u r c h g e k e n n z e i c h n e t,  
dass der Text von Dokumenten in grafischer Darstellung  
25 zunächst mit einem OCR-Verfahren in codierten Text umgesetzt  
wird und die Vorschlags-Routine zusätzlich zum markierten  
Stringabschnitt in grafischer Darstellung den codierten Text  
dieses Stringabschnittes darstellt.
- 30 9. Verfahren nach einem der Ansprüche 1 bis 7,  
d a d u r c h g e k e n n z e i c h n e t,  
dass zusätzlich zum markierten Stringabschnitt dieser  
Stringabschnitt nochmals in vergrößerter Darstellung am  
Bildschirm angezeigt wird.

10. Verfahren nach einem der Ansprüche 1 bis 9,  
dadurch gekennzeichnet,  
dass die Vorschlags-Routine nach dem Markieren eines  
Stringabschnittes eine Funktion aktiviert, mit welcher  
5 durch Betätigen einer oder mehrerer vorbestimmter Tasten  
der Inhalt des markierten Stringabschnittes in die Daten-  
bank übernommen wird.
11. Verfahren nach einem der Ansprüche 1 bis 10,  
10 dadurch gekennzeichnet,  
dass beim Ausführen der Vorschlags-Routine nach dem Bewe-  
gen des Zeigers ein vorbestimmtes Zeitintervall abgewart-  
tet wird, innerhalb dessen der Zeiger nicht bewegt werden  
darf, bevor ein Stringabschnitt ausgewählt wird.
12. Verfahren zum Erfassen von Daten aus maschinell lesbaren  
15 Dokumenten, wobei die Daten einer Datenbank zugeordnet  
werden, insbesondere nach einem der Ansprüche 1 - 11,  
dadurch gekennzeichnet,  
20 dass nach dem Einlesen von Daten aus einer ersten Tabel-  
lenzeile in korrespondierende Datenbankfelder automatisch  
die weiteren Tabelleneinträge durch einen Vergleich von  
unterhalb der letzten Tabellenzeile angeordneten String-  
abschnitten mit den Stringabschnitten der ersten Tabel-  
25 lenzeile, die den eingelesenen Daten entsprechen, ermit-  
telt werden und diese weiteren Tabelleneinträge automa-  
tisch extrahiert werden.
13. Verfahren nach Anspruch 12,  
30 dadurch gekennzeichnet,  
dass der Vergleich zwischen den Stringabschnitten mittels  
eines String Matching-Verfahrens erfolgt.
14. Verfahren nach Anspruch 12 oder 13,  
35 dadurch gekennzeichnet,  
dass die ermittelten Tabelleneinträge markiert werden.

15. Verfahren nach Anspruch 14,  
dadurch gekennzeichnet,  
dass Funktionen zum Editieren der markierten Tabellenein-  
träge bereitgestellt werden.
- 5
16. System zum Erfassen von Daten aus maschinell lesbaren Do-  
kumenten umfassend einen Computer (12) mit einer Spei-  
chereinrichtung (13) und eine CPU (14), wobei in der  
Speichereinrichtung (13) ein Softwareprodukt zum Ausfüh-  
10 ren des Verfahrens nach einem der Ansprüche 1 - 15 ge-  
speichert ist.
17. System nach Anspruch 16,  
dadurch gekennzeichnet,  
15 dass das System eine Eingabeeinrichtung (17) in Form ei-  
ner Maus (6) und/oder Tastatur (7) aufweist.
18. System nach Anspruch 16 oder 17,  
dadurch gekennzeichnet,  
20 dass das System einen Scanner (16) zum optischen Abtasten  
von Dokumenten aufweist.
19. Computerprogrammprodukt, das bei seinem Laden und Ausfüh-  
ren auf einem Computer (12) ein Verfahren nach einem der  
25 Ansprüche 1 bis 15 bewirkt.

1/8

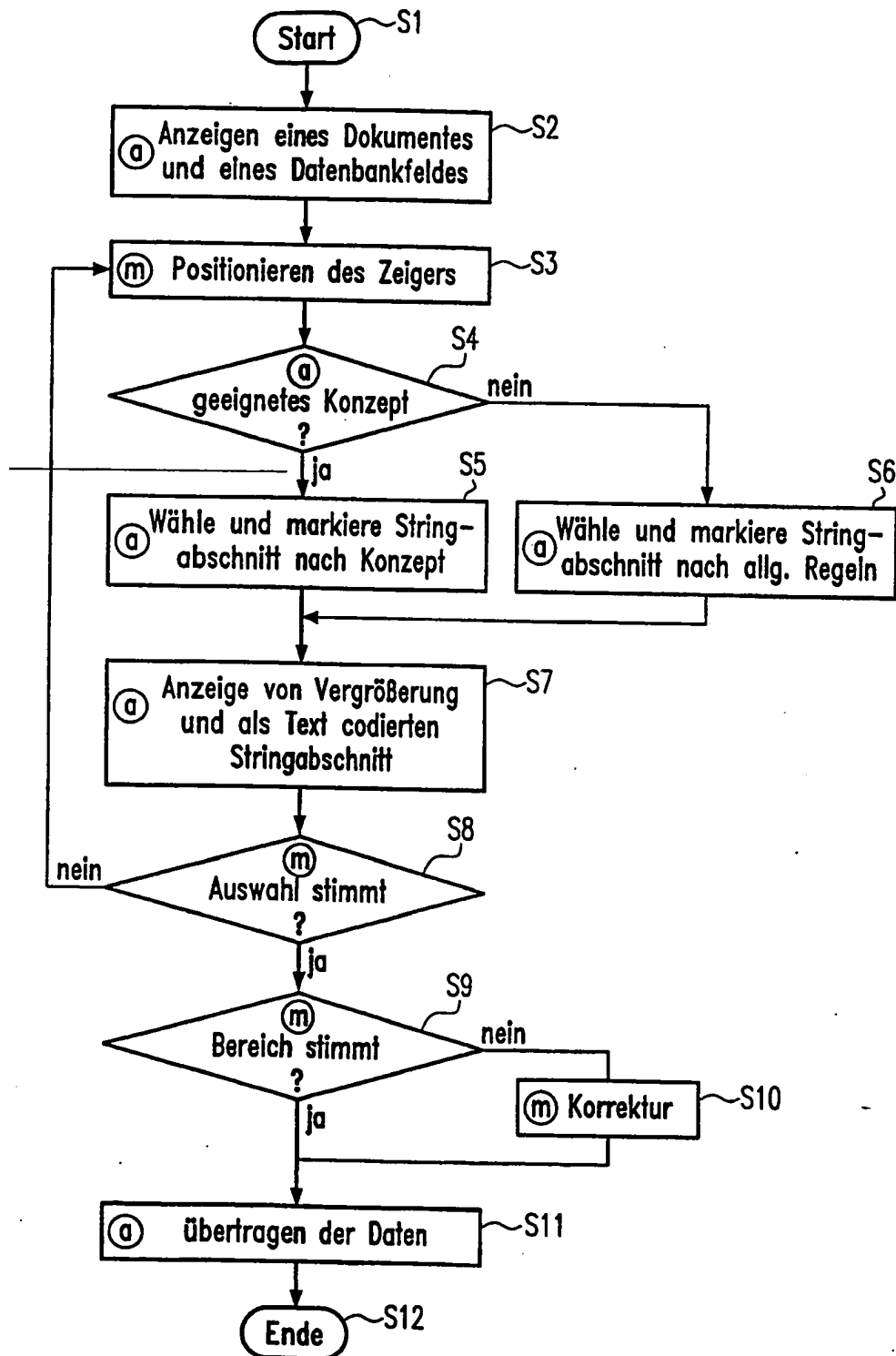
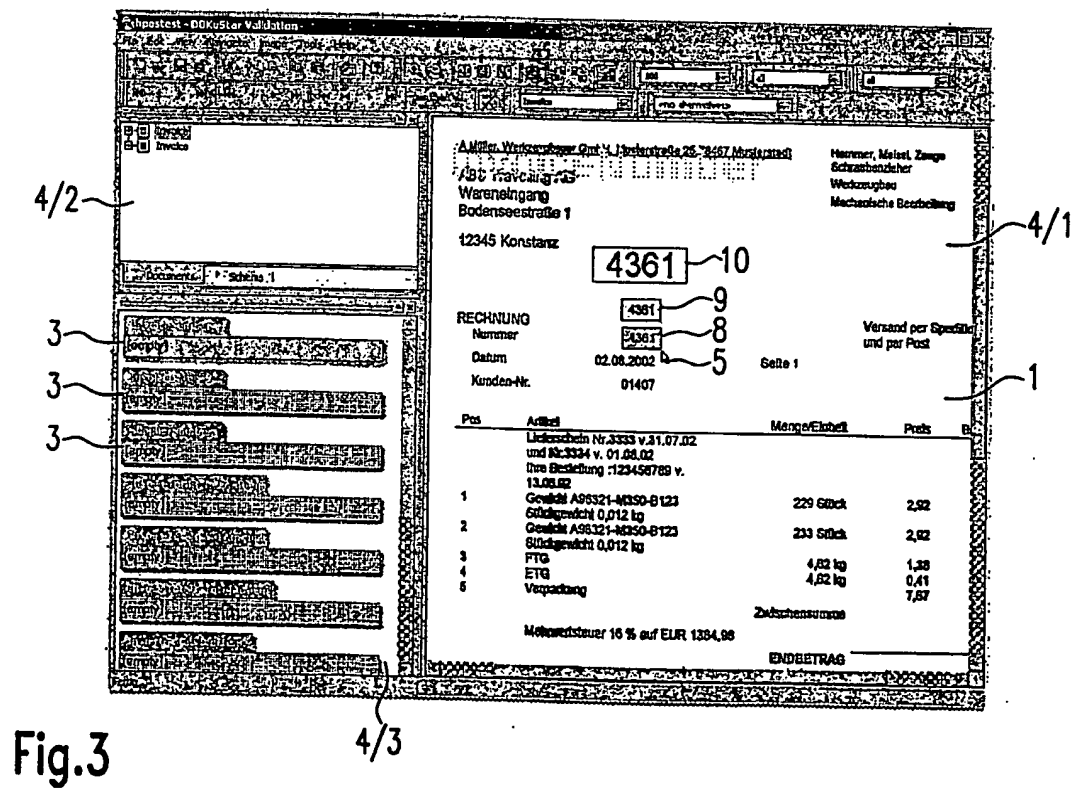
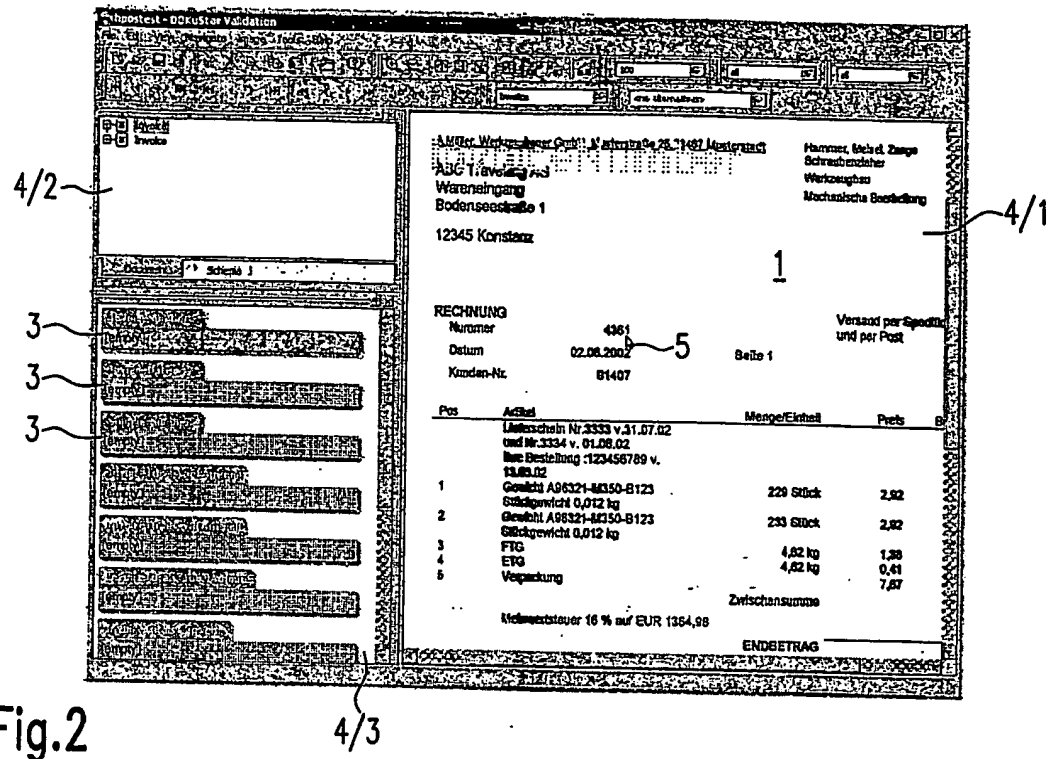


Fig.1

2/8





3/8

4/2

3

3

3

4/1

1

4/3

AMMEX Werkzeugfabrik GmbH Musterstraße 72, 79567 Musterstadt  
 12345 Konstanz

Hammer, Meißel, Zange  
 Schneidmesser  
 Werkzeugkasten  
 Mechanische Bearbeitung

4361 10  
 4361 9  
 4361 8  
 4361 5

RECHNUNG  
 Nummer  
 Datum  
 Kunden-Nr.

02.08.2002  
 01487

Seite 1

Versand per Spedition  
 und per Post

Pos	Artikel	Menge/Einheit	Preis	Br.
1	Lieferanten Nr. 3333 v. 31.07.02 und Nr. 3334 v. 01.08.02 Ihre Bestellung: 123456789 v. 13.08.02 Gewicht A98321-M350-S123 Stückgewicht 0,012 kg	229 Stück	2,92	
2	Gewicht A98321-M350-S123 Stückgewicht 0,012 kg	233 Stück	2,92	
3	FTG	4,82 kg	1,38	
4	ETG	4,82 kg	0,41	
5	Verpackung		7,67	
Zwischensumme				
Mehrwertsteuer 15 % auf EUR 1384,98				
ENDBETRAG				

Fig.4

4/2

3

3

3

4/1

1

4/3

AMMEX Werkzeugfabrik GmbH Musterstraße 72, 79567 Musterstadt  
 12345 Konstanz

Hammer, Meißel, Zange  
 Schneidmesser  
 Werkzeugkasten  
 Mechanische Bearbeitung

4361 10  
 4361 9  
 4361 8  
 4361 5

RECHNUNG  
 Nummer  
 Datum  
 Kunden-Nr.

02.08.2002  
 01487

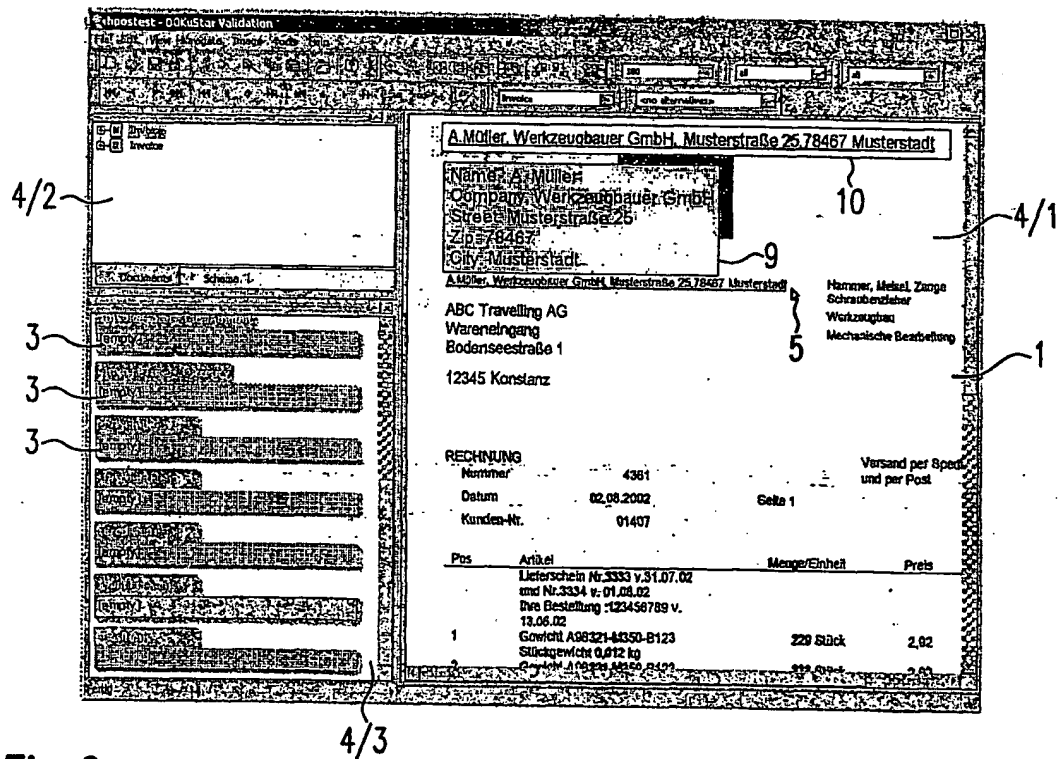
Seite 1

Versand per Spedition  
 und per Post

Pos	Artikel	Menge/Einheit	Preis	Br.
1	Lieferanten Nr. 3333 v. 31.07.02 und Nr. 3334 v. 01.08.02 Ihre Bestellung: 123456789 v. 13.08.02 Gewicht A98321-M350-S123 Stückgewicht 0,012 kg	229 Stück	2,92	
2	Gewicht A98321-M350-S123 Stückgewicht 0,012 kg	233 Stück	2,92	
3	FTG	4,82 kg	1,38	
4	ETG	4,82 kg	0,41	
5	Verpackung		7,67	
Zwischensumme				
Mehrwertsteuer 15 % auf EUR 1384,98				
ENDBETRAG				

Fig.5

4/8



**Fig.6**

5/8

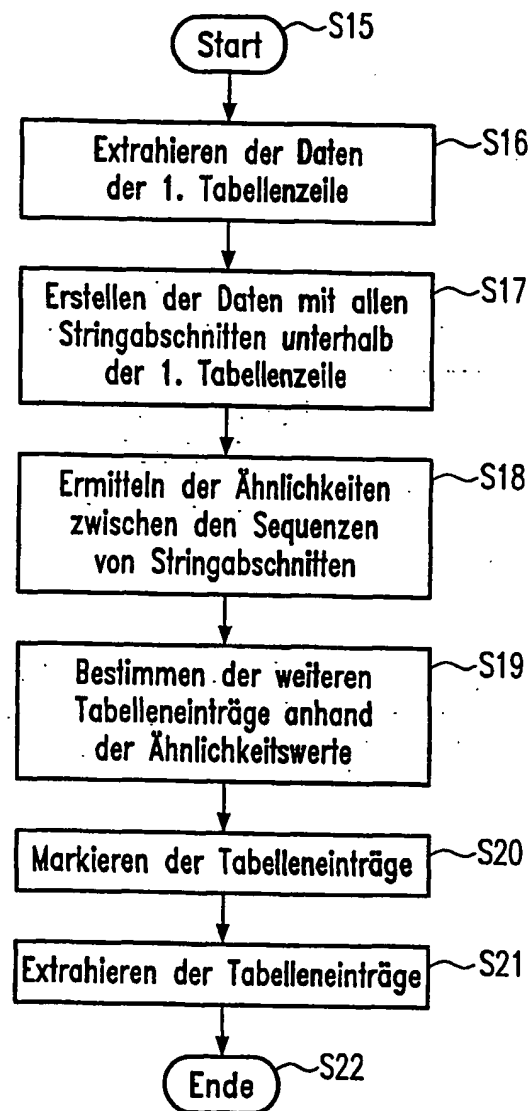


Fig.7

6/8

18. Januar 2003 Seite 1

Maschine: [REDACTED]

Geschäftsbeziehung: [REDACTED]

Durchschnitt: [REDACTED]

Elektrische: [REDACTED]

Geräte: [REDACTED]

30173 Hymenopter

Bestellung (geändert)  
Bestell-Nr.: 31335-200203

Menge	Bezeichnung der Leistung	Liefertermin	Einheitspreis	Betrag
Wir bestellen zu den Bedingungen des HGB/BGB:				
2.000 kg	Unsere Artikel-Nr. 1011 Ultraschall A3-WG 7, natur	20.01.03		
3.000 kg	Unsere Artikel-Nr. 1187 Ultraschall A3-K, natur	14.01.03	2,40 EUR	7.200,00
1.000 kg	Unsere Artikel-Nr. 1770 Ultraschall A4H	20.01.03		
3.000 kg	Unsere Artikel-Nr. 1015 Ultraschall B3-WG 7, natur	20.01.03		
8.000 kg	Unsere Artikel-Nr. 1014 Ultraschall B3-WG 6, natur	20.01.03		
5.000 kg	Unsere Artikel-Nr. 1506 Ultraschall B3-ZG 6 schwarz	20.01.03		
6 Stück	Unsere Artikel-Nr. 200028 Webstanzgerät nach DIN 50049-2.2	20.01.03		
1.000 kg	Unsere Artikel-Nr. 1032/207478 1 Ultraschall B3-WG 6, schwarz	20.01.03		
1 Stück	Unsere Artikel-Nr. 200019 Webstanzgerät nach DIN 50049-2.2	20.01.03		
				7.200,00

Zahlbar innerhalb 14 Tage mit 3 % Skonto, 30 Tage netto.  
Lieferung erfolgt frei Haus, einschließlich Verpackung.

Fig.8

7/8

16. Januar 2003 Seite 1

Kunde: [REDACTED]

Stichtag: [REDACTED]

Durchgeführt: [REDACTED]

Erstellt: [REDACTED]

30173 Hannover

Bestellung (geändert)  
Bestell-Nr.: 31335-200303

Menge	Beschreibung der Leistung	Liefertermin	Einzelpreis	Summe
Wir bestellen zu den Bedingungen des KVB/BGB:				
2.000 kg	Unsere Artikel-Nr. 1011 Ultramid A3 WS 7, natur	20.01.03		
8.000 kg	Unsere Artikel-Nr. 1197 Ultramid A3 K, natur	14.01.03	2,40 EUR	7.200,00
1.000 kg	Unsere Artikel-Nr. 1770 Ultramid A4H	20.01.03		
3.000 kg	Unsere Artikel-Nr. 1015 Ultramid B3 WS 7, natur	20.01.03		
3.000 kg	Unsere Artikel-Nr. 1014 Ultramid B3 WS 5, natur	20.01.03		
3.000 kg	Unsere Artikel-Nr. 1505 Ultramid B3 ZB 6 schwarz	20.01.03		
10 Stk	Unsere Artikel-Nr. 200018 Werkzeuggüte nach DIN 50049-2.2	20.01.03		
1.000 kg	Unsere Artikel-Nr. 1032/201478.1 Ultramid B3 ZB 6 schwarz	20.01.03		
1 Stk	Unsere Artikel-Nr. 200019 Werkzeuggüte nach DIN 50049-2.2	20.01.03		
			<b>Summe</b>	<b>7.200,00</b>

Zahlung innerhalb 14 Tage mit 3 % Skonto, 30 Tage netto.  
Lieferung erfolgt frei Haus, einschließlich Verpackung.

Fig.9

8/8

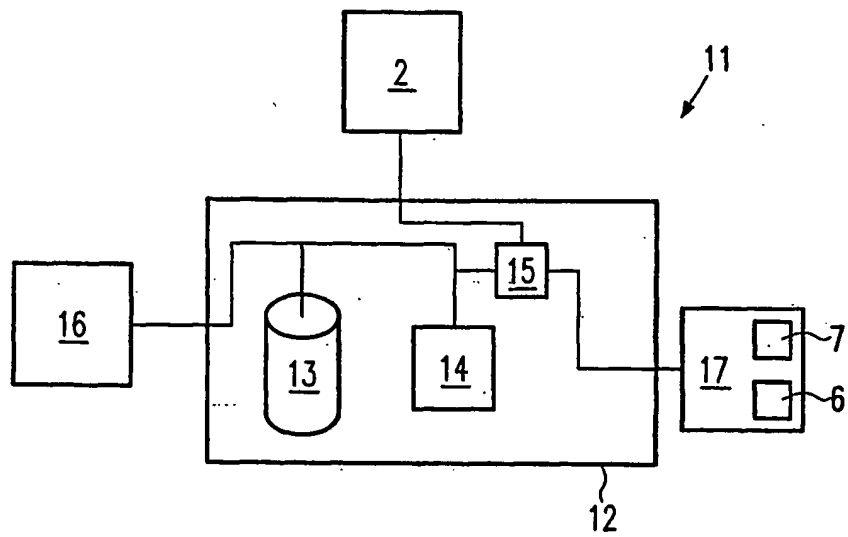


Fig.10

**This Page is Inserted by IFW Indexing and Scanning  
Operations and is not part of the Official Record**

**BEST AVAILABLE IMAGES**

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ BLACK BORDERS
- ☐ IMAGE CUT OFF AT TOP, BOTTOM OR SIDES
- ☐ FADED TEXT OR DRAWING
- ☒ BLURRED OR ILLEGIBLE TEXT OR DRAWING
- ☐ SKEWED/SLANTED IMAGES
- ☐ COLOR OR BLACK AND WHITE PHOTOGRAPHS
- ☐ GRAY SCALE DOCUMENTS
- ☐ LINES OR MARKS ON ORIGINAL DOCUMENT
- ☐ REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY
- ☐ OTHER: \_\_\_\_\_

**IMAGES ARE BEST AVAILABLE COPY.**

**As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.**